responding to the present Office Action and that the requirements for applications containing nucleotide and/or amino acid sequences be met.

The Examiner objected to claim 5 regarding an alleged informality and to the disclosure for having an embedded hyperlink appearing therein. The Examiner rejected claim 4 under 35 U.S.C. §112, first paragraph, for an alleged lack of a supporting enabling disclosure and for allegedly containing subject matter which is not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention. The Examiner rejected claims 1-12 under 35 U.S.C. §112, second paragraph, as allegedly being indefinite for failing to particularly point out and distinctly claim the invention.

The Examiner rejected claims 1-8, 10-12, 23 and 25 under 35 U.S.C. §102(b) as allegedly being unpatentable over the article entitled "GenBank" by Benson et al. (hereinafter "Benson").

The present invention is directed to a method, system and an article of manufacture wherein a set of sequences is provided. The sequences are not aligned. A plurality of patterns common to a plurality of the sequences are discovered. It is determined if a candidate sequence comprises a predetermined number of the patterns.

Applicants have amended the specification to correct typographical and grammatical errors appearing therein. Specifically, on page 4, line 11 of the specification, "widows" has been replaced with the corrected "windows." On page 13, line 8 of the specification, the article "a" has been added between "is" and "good choice." On page 26, line 25 of the specification, the word "do" has been deleted. On page 31, line 14 of the specification, "sequence" has been replaced with the correct plural form "sequences." On page 31, line 20 of the specification, "sud-selected" has been replaced with "sub-selected" to correct the misspelling.

ELECTION/RESTRICTION REQUIREMENT

The Examiner contended that the group containing claims 1-12, 23 and 25 is directed to a method of discovering a plurality of patterns and determining if a candidate sequence is comprised of a predetermined number of patterns; and that the group containing claims 13-22, 24 and 26 is directed to a method for unsupervised building and exploitation of composite descriptors. The Examiner alleged that the distinct critical features of each group support an undue search burden if they were examined together, and as such, restriction of the groups is proper.

In a telephone interview on April 23, 2002, Applicants elected claims 1-12, 23 and 25, without traverse, for prosecution on the merits. Applicants would like to thank Examiner Sheinberg for conducting the interview. Accordingly, Applicants herein affirm the election of claims 1-12, 23 and 25, without traverse.

DRAWING CHANGES

In the Office Action, the Examiner noted that the required timing for drawing corrections has changed. As such, the Examiner has required that any corrections to the drawings in the instant application be submitted within the time period set for responding to the present Office Action. In compliance with the timing requirements for submitting corrections to the drawings, a copy of the formal drawings are being submitted herewith.

SEQUENCE COMPLIANCE

The Examiner highlighted that the instant application contains sequence disclosures that are encompassed by the definition of amino acid sequences, as set forth in 37 C.F.R. §1.821(a)(1) and (a)(2). Therefore, the Examiner required compliance with the rules governing the presentation of these amino acid sequences, i.e., as outlined in 37 C.F.R. §1.821 through 1.825. In compliance with those requirements, Applicants submit herewith a written sequence listing and an identical copy of that sequence listing in a computer-readable form. Applicants attest that the information recorded in the computer-readable form is identical to the written sequence listing.

Further, Applicants have amended the specification to properly reference the sequence listing. Specifically, Applicants have amended the information appearing in Table 1 on page 28, Tables 2 and 3 on page 29, Table 4 on page 30 and the sequences presented under the heading "Helix-Turn-Helix" on page 48, lines 5-14. The phrase "SEQ ID NO" has been added before each sequence followed by a numerical designation that corresponds to that particular sequence in the sequence listing, e.g., the phrase "SEQ ID NO 1" has been added before the first sequence presented in Table 1 to indicate that the sequence that follows is SEQ ID NO 1 in the sequence listing.

CLAIM OBJECTIONS

The Examiner objected to claim 5 due to an alleged informality appearing therein. Namely, the Examiner objected to the absence of a comma between the terms "of" and "if" appearing on line 1 of claim 5. Applicants have amended claim 5 to clarify the language appearing therein. A

colon has been added after "of" in line 1 of claim 5, and the claim steps have been separated by a semi colon.

The Examiner further objected to the disclosure as having an embedded hyperlink. To address the Examiner's concerns, the specification has been amended at page 43, lines 10-14, to remove the hyperlink and in its place include the hyperlink source, namely, the Washington University in St. Louis, School of Medicine, Genome Sequence Center.

FORMAL REJECTIONS

As previously indicated, the Examiner rejected claim 4 under 35 U.S.C. § 112, first paragraph, for allegedly lacking a supporting enabling disclosure. Specifically, the Examiner contended that, while the method of claim 4 is supported by the specification when the properties and features of the sequences are known, e.g., for EF1G/PS50040, enablement is not provided for the method being performed without any knowledge of the properties or features of the sequence, as in claim 4. Applicants respectfully disagree.

The specification clearly enables one of ordinary skill in the art to discover a plurality of patterns common to a plurality of sequences without any knowledge of properties or features of the sequences, as in claim 4. By way of example only, the second example on page 30, line 4, through page 31, line 24 of the specification illustrates the present method utilizing a generalized set of sequences. To obtain a generalized set of sequences, an original set (from the GPCRDB as of May 1998) was intersected with an older release of Swiss-Prot, i.e., Release 35.0 from November, 1997. See page 31, lines 4-5 of the specification. The intersected sets were then randomized. See page 31, lines 13-14 of the specification. A training set was formed by selecting the sequences and fragments in the first 80 positions. Since the sets were randomized, selecting the first 80 positions was a random selection. See page 31, lines 17-24 of the specification.

Applicants respectfully submit that the above teachings do not utilize any known properties or features of the sequences in the original sets and would clearly enable one of ordinary skill in the art to perform the techniques of the present invention on any given set of sequences. The Examiner has pointed out that the teachings highlighted in the specification use sequence sets with known features, such as the GPCR entries described herein. However, sequence sets, such as the GPCR entries, are used merely as exemplary data sets and the properties thereof are not needed to perform the instant techniques, as is clearly illustrated by the above example.

As previously indicated, the Examiner also rejected claim 4 under 35 U.S.C. § 112, first paragraph, as containing subject matter which is not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention. Specifically, the Examiner contended that the specification lacks sufficient description of a sequence wherein there is not any knowledge of the properties or features of the sequence, as in claim 4, because a sequence in itself is a specific property or feature of the sequence. Applicants respectfully disagree with the Examiner's assertions.

Claim 4 recites that a plurality of patterns common to a plurality of sequences are discovered without using any knowledge about properties or features of the sequences. The instant specification clearly supports discovering a plurality of patterns common to a plurality of sequences without using any knowledge about properties or features of the sequences. By way of example only, the second example, highlighted immediately above, and directed to sequences taken from the GPCRDB, does not employ any information regarding the properties and/or features of the sequences contained therein to perform the instant method. Even if it is assumed that the Examiner's argument is correct, and a sequence in and of itself is a specific property of the sequence, neither that property nor any other property of the sequence is used in the example. As such, the method, as exemplified in claim 4, may be performed without using any knowledge about the properties and/or features of the sequences.

As previously indicated, the Examiner also rejected claims 1-12 under 35 U.S.C. § 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the invention. First, the Examiner contended that the references in claims 1, 4, 23 and 25 to sequences that are "not aligned," or in the case of claim 4 "unaligned," renders the claims vague and indefinite. As such, claims 2, 3 and 5-12 are also rejected as being dependent from claim 1. The Examiner alleged that it is unclear what criteria are being used to determine that a set of sequences is not aligned.

Applicants would like to direct the Examiner's attention to page 10, lines 1-13 of the specification, wherein a process for aligning sequences is described. As highlighted in an example given therein, aligning patterns within sequences, e.g., sequences: {DEFXYZABC} and {ABCDEF} and patterns: ABC and DEF, can be problematic because there isn't a single, unique alignment. Basically, either the pattern ABC can be aligned in the two sequences, or alternatively, the pattern DEF can be aligned in the two sequences, but not both. In the present invention, no alignment of sequences is required. Thus, according to the above example, both of the sequences {DEFXYZABC} and {ABCDEF} may be searched for either pattern ABC, DEF, or both, and each of the patterns would be found as common to the two sequences.

Second, the Examiner contended that the reference in claims 1, 2 and 5 to the phrases "candidate sequence" and "patterns" renders the claims vague and indefinite. As such, claims 3, 4 and 6-12 are also rejected as being dependent from claim 1. The Examiner alleged that it is unclear what criteria are being used to determine that a sequence is a candidate sequence.

Applicants respectfully submit that, given the description provided in the specification and knowledge common to those of ordinary skill in the art, use of the terms "candidate sequence" and "patterns" do not render the claims vague and indefinite. By way of example only, the specification provides that patterns common to some, or all, of the sequences in a set of unaligned sequences are discovered. The patterns are then used to determine if a candidate sequence is a member of that set, i.e., family, of sequences. See page 6, lines 13-15 of the specification. Thus, the candidate sequence is a sequence that may or may not be included as a member of the family of sequences. Whether or not the candidate sequence is included as a member of the family of sequences may be determined using patterns common to some or all of the sequences in the set. Therefore, the use of the terms "candidate sequence" and "patterns" is not unclear.

Third, the Examiner contended that the phrase "some of the positions each comprising at least one expected symbol and other of the positions comprising "don't care" positions," in claim 6, renders the claim vague and indefinite due to improper English grammar. Further, the Examiner contended that the phrase "don't care," appearing therein, is also vague and indefinite, as it is unclear what criteria are being used to determine "don't care" positions. As such, claim 7 is also rejected as being dependent from claim 6.

To address the Examiner's concerns, Applicants have amended claim 6 to clarify the claim wording, as well as to remove the phrase "don't care" and add the equivalent "positions which may be occupied by any sequence character." Thus, claim 6 recites, in part, that each pattern, of claim 1, comprises a plurality of positions. Some of the plurality of positions each comprise at least one expected symbol and other of the plurality of positions comprise positions which may be occupied by any sequence character. As such, Applicants respectfully request reconsideration of the claim.

Applicants however point out that the phrase "don't care," given the description provided in the specification and knowledge common to those of ordinary skill in the art, is not vague and indefinite. The term "don't care" position refers to positions which may be occupied by any sequence character. Pattern discovery, including "don't care" positions, is described, for example, in Floratos, et al., U.S. Patent No. 6,108,666, "Method and Apparatus for Pattern Discovery in 1-Dimensional Systems," which has been incorporated by reference in the instant application. Thus, for example, if

sequences comprise characters of the English alphabet, then any one of the 26 letters of the English alphabet may be used to fill a "don't care" position in a related pattern.

PRIOR ART REJECTION

As previously indicated, the Examiner rejected claims 1-8, 10-12, 23 and 25 under 35 U.S.C. §102(b) as allegedly unpatentable over Benson. Applicants respectfully disagree with the Examiner's rejection.

Benson is directed to a description of the GenBank sequence database. Entrez, an integrated retrieval system of GenBank, serves to incorporate data from major DNA and protein sequence databases, as well as genome maps and protein structure information. Sequence similarity searching may be achieved using the BLAST family of programs. See article abstract.

The Examiner contended that Benson discloses GenBank which contains 602,072,354 nucleotides from 920,588 different sequences, as in the instant claims 4 and 6. Further, that human entries in the primate division of GenBank are combined with human expressed sequence tags (ESTs) to create a cluster of sequences that share virtually identical 3' untranslated regions, as in the instant claims 2 and 7. Further, that GenBank may be used for sequence similarity searching. Finally, that the combination of these disclosures also anticipate the instant claim 1. Applicants respectfully disagree.

Applicants find no teaching or suggestion in Benson of either discovering a plurality of patterns common to a plurality of sequences in a set of sequences, and/or determining if a candidate sequence comprises a predetermined number of the patterns, both limitations present in Applicants' independent claims 1, 23 and 25. Essentially, Benson merely discloses the GenBank sequence database. The Examiner highlighted that Benson mentions that the BLAST program may be used to search the database. Even so, the use of BLAST to search GenBank does not teach or suggest the present invention.

The BLAST program functions by taking a query sequence, generating *k*-tuples from the query sequence, wherein the value of *k* is fixed, and searching for instances of the *k*-tuples in a database. As such, BLAST does not teach or suggest generating a plurality of patterns common to a plurality of sequences in a set of sequences, as is required in independent claims 1, 23 and 25. At most, use of the BLAST program might generate duplicate *k*-tuples from a particular query sequence, which is in no way indicative of any patterns present in the set of sequences.

Further, BLAST does not teach or suggest first discovering a plurality of patterns common to a plurality of sequences in a set of sequences, and then determining if a candidate sequence

comprises a predetermined number of the patterns, as is also required in independent claims 1, 23 and 25. As mentioned above, BLAST does not teach generating a plurality of patterns common to a set of sequences. BLAST also does not compare anything, pattern or otherwise, to the candidate sequence. Further mentioned above, BLAST generates *k*-tuples from the query sequence, and as such, the query sequence is the candidate sequence. The *k*-tuples are then compared to the database sequences. As such, for at least the above-stated reasons, Benson does not anticipate, nor make obvious, the teachings of the present invention.

Since the teachings of Benson do not disclose all limitations of Applicants' independent claims 1, 23 and 25, such claims are allowable. As claims depending from an allowable claim, dependent claims 2-8, 10-12 are also allowable. However, Applicants also assert that said dependent claims recite patentable subject matter in their own right. Therefore, Applicants respectfully request reconsideration and withdrawal of the rejections, and allowance of the claims.

In view of the foregoing, the invention, as claimed in claims 1-8, 10-12, 23 and 25, cannot be said to be either taught or suggested by Benson. Accordingly, Applicants submit all of the pending claims, i.e., claims 1-12, 23 and 25, are in condition for allowance and such favorable action is earnestly solicited.

If any outstanding issues remain, or if the Examiner has any further suggestions for expediting allowance of this application, the Examiner is invited to contact the undersigned at the telephone number indicated below.

The Examiner's attention to this matter is appreciated.

Attached hereto is a marked-up version of changes made to the specification and claims by the present Amendment.

Respectfully submitted,

Dated: May 22, 2003

Kevin M. Mason
Attorney for Applicant(s)
Reg. No. 36,597
Ryan, Mason & Lewis, LLP
1300 Post Road, Suite 205
Fairfield, CT 06824
(203) 255-6560

VERSION WITH MARKINGS TO SHOW CHANGES MADE

IN THE SPECIFICATION:

Please amend the paragraph as it appears on page 3, line 27, through page 4, line 14 as follows:

In addition to descriptor approaches, there are also "windowing" approaches that build descriptors for a family. In these methods, one or more windows are used instead of character patterns. A single window method is called the PROFILE approach. All of the sequences of each of the family members are aligned with respect to their best-conserved region. Researchers then determined a probability distribution for locations in each column of the implied window. For each such block, they determined a probability of expecting an amino acid at some location within the window and thus built a 'profile' of expected probabilities for each of the columns of the window. The researchers would slide this set of probabilities against an unknown protein. If this candidate protein matched the expected probabilities, they included the protein as a member of the family. This approach was more tolerant than the single descriptor approach. Subsequently, researchers began to use profiles for multiple [widows]windows. There could be two, three, four windows where the members of the family could agree on content. Sometimes, a profile was not built explicitly but rather was maintained as a collection of the instances across the known or alleged family members of the conserved region under consideration.

Please amend the paragraph as it appears on page 13, lines 5-12 as follows:

In step 210, the sequence threshold, K, is set. It is possible to set K=|T|, which is the number of sequences in the training set. In actuality, it has proven beneficial to assign a small starting value to K that is a fraction of the number of sequences in T. Experiments have shown that a starting value of K=|T|/b with b=4 or 5 is a good choice across many data sets. Note that the smaller the value of b, the higher the redundancy of the composite descriptor will be. The selection of K also can depend on how conserved, or similar, the family members are. If the family members are well conserved, then K can be higher; if the family members are not well conserved, then K can be lower.

Please amend the paragraph as it appears on page 26, line 23, through page 27, line 5 as follows:

Note that the 5 hits SYV_FUGRU, GTT1_RAT, GTT1_MOUSE, SYEP_HUMAN and GTH4_MAIZE are clearly separated from the 11 top scoring sequences. They [do] however obtained scores which were above threshold and thus are studied in more detail. In all 5 cases, one or more sizeable regions that were shared with one or more members of the PS50040 collection were discovered. The Clustal-W alignment of EF1G_XENLA and the N-terminus of SYV_FUGRU, a valyl-trna synthetase from Fugu rubripes, are shown in Table 1 below. Table 1 shows a Clustal-W alignment of EF1G_XENLA and the N-terminus of SYV_FUGRU, and this shows a strong similarity. As can be seen, the similarity among these two sequences is pretty extended and the Clustal-W score for the shown alignment equaled 462.

Please amend Table 1 as it appears on page 28 as follows:

Table 1

```
EF1G_XENLA       (SEQ ID NO 1) MAGGTLYTYPDNWRAYKPLIAAQYSGFPIKVASSAPEFQFGVTNKTPEFLKKFPLGKVPA
SYV_FUGRU_piece  (SEQ ID NO 2) MA--TLYVSP-----------HLDDFRSLLALVAAEY---------------------
                                ** ***. *             : ..*   :*  *.*:

EF1G_XENLA       (SEQ ID NO 3) FEGKDGFCLFESSAIAHYVGNDELRGTTRLHQAQVIQWVSFSDSHIVPPASAWVFPTLGI
SYV_FUGRU_piece  (SEQ ID NO 4) -------C----------GNAKQ-------QSQVWQWLSFADNELTPVSCAVVFPLMGM
                                       *          ** :        *:** **:**:*..:.* :.* *** :*:

EF1G_XENLA       (SEQ ID NO 5) MQYNKQATEQAKEGIKTVLGVLDSHLQTRTFLVGERITLADITVTCSLLWLYKQVLEPSF
SYV_FUGRU_piece  (SEQ ID NO 6) TGLDKKIQQNSRVELMRVLKVLDQALEPRTFLVGESITLADMAVAMAVLLPFKYVLEPSD
                                  :*:  :::: :  ** ***. *:.******* *****::*: ::*  :* *****

EF1G_XENLA       (SEQ ID NO 7) RQPFGNVTRWFVTCVNQPEFRAVLGEVKLCDKMAQFDAKKFAEMQPKKETPKKEKPAKEP
SYV_FUGRU_piece  (SEQ ID NO 8) RNVLMNVTRWFTTCINQPEFLKVLGKISLCEKMVPVTAKTSTEEAAAVH-PDAAALNGPP
                               *: : ******.**:*****  ***::.**:**. . **. :*  .  . *.       *

EF1G_XENLA       (SEQ ID NO 9) KKEKEEKKKAAPTPAPAPEDDLDESEKALAAEPKSKDPYAHLP-KSSFIMDEFKRKYSNE
SYV_FUGRU_piece  (SEQ ID NO 10)KTEAQLKKEAKKREKLEKFQQKKEMEAKKKMQPVAEKKAKPEKRELGVITYDIPTPSGEK
                               *.* : **:*         :: .* *     :* ::.       : ..*  ::    .::

EF1G_XENLA       (SEQ ID NO 11)DTLTVALPYFW-EHFDKEGWSIWYAEY-KFPEELTQAFMSCNLITGMFQR-LDKLRKTGF
SYV_FUGRU_piece  (SEQ ID NO 12)KDVVSPLPDSYSPQYVEAAWYPWWEKQGFFKPEFGRKSIGEQNPRGIFMMCIPPPNVTGS
                               . :. .**  :  :: : .*  *: :   *  *: :  :. :   *:*   :   . **

EF1G_XENLA       (SEQ ID NO 13)ASVILFGTNNNSSISGVWV-FRGQDLAFTLSED-----WQIDYESYNWRKLDSGSEEC--
SYV_FUGRU_piece  (SEQ ID NO 14)LHLGHALTNAIQDTLTRWHRMRGETTLWNPGCDHAGIATQVVVEKKLMREKGTSRHDLGR
                                :    **  ..    *  :**:   :. . *      *:  *.   *: .:. .:

EF1G_XENLA       (SEQ ID NO 15)KTLVKEYFAWEGE--------FKNVGKPFNQG-KIFK----------------------
SYV_FUGRU_piece  (SEQ ID NO 16)EKFIEEVWKWKNEKGDRIYHQLKKLGSSLDWDRACFTMDPKLSYAVQEAFIRMHDEGVIY
                               :.:::* : *:.*        :*::*..:: .   *.
```

Please amend Table 2 as it appears on page 29 as follows:

Table 2

```
GTT1_MOUSE (SEQ ID NO 17) -VLELYLDLLSQPCRAIYIFAKKNNIPFQMHTVELRKGEHLSDAFARVNPMKKVPAMM-D
GTT1_RAT   (SEQ ID NO 18) -VLELYLDLLSQPCRAIYIFAKKNNIPFQMHTVELRKGEHLSDAFAQVNPMKKVPAMK-D
EF1G_ARTSA (SEQ ID NO 19) VAGKLYTYPENFRAFKALIAAQYSGAKLEIAKSFVFGETNKSDAFLKSFPLGKVPAFESA
                           . :**    . .    * *: ..  ::: .  :    : **** :  *: ****:

GTT1_MOUSE (SEQ ID NO 20) GGFTLCESVAILLYLAHK-----------------------YKVPDHWYPQDLQARARV
GTT1_RAT   (SEQ ID NO 21) GGFTLCESVAILLYLAHK-----------------------YKVPDHWYPQDLQARARV
EF1G_ARTSA (SEQ ID NO 22) DGHCIAESNAIAYYVANETLRGSSDLEKAQIIQWMTFADTEILPASCTWVFPVLGIMQFN
                           .*. :.** **  *:*::                          ..  *    *

GTT1_MOUSE (SEQ ID NO 23) DEYLAWQHTGLRRSCLRALWHKVMFPVFLGEQIPPETLAATLAELDVNLQVLEDKFLQDK
GTT1_RAT   (SEQ ID NO 24) DEYLAWQHTTLRRSCLRTLWHKVMFPVFLGEQIRPEMLAATLADLDVNVQVLEDQFLQDK
EF1G_ARTSA (SEQ ID NO 25) KQATARAKEDIDKALQALDDHLLTRTYLVGERITLADIVVTCTLLHLYQHVLDEAFRKSY
                           .:  *  :  : ::      * :  . ::**:*    :..* : *.:  :**:: * :.

GTT1_MOUSE (SEQ ID NO 26) DFLVGPHISLADLVAITELMHPVGGGCPVFEGHPRLAAWYQRVEAAVGKDLFREAHEVIL
GTT1_RAT   (SEQ ID NO 27) DFLVGPHISLADVVAITELMHPVGGGCPVFEGRPRLAAWYRRVEAAVGKDLFLEAHEVIL
EF1G_ARTSA (SEQ ID NO 28) VNTNRWFITLINQKQVKAVIGDFKLCEKAGEFDP---KKYAEFQAAIGSGEKKKTEKAPK
                                .*:* :   :. ::  .     . *  *     * ..:**:*..   ::.:.

GTT1_MOUSE (SEQ ID NO 29) KVKDCPPADLIIKQKLMPRVLTMIQ----------------------------------
GTT1_RAT   (SEQ ID NO 30) KVRDCPPADPVIKQKLMPRVLTMIQ----------------------------------
EF1G_ARTSA (SEQ ID NO 31) AVKAKPEKKEVPKKEQEEPADAAEEALAAEPKSKDPFDEMPKGTFNMDDFKRFYSNNEET
                           *:  *  . : *::    . :  :

GTT1_MOUSE (SEQ ID NO 32) ------------------------------------------------------------
GTT1_RAT   (SEQ ID NO 33) ------------------------------------------------------------
EF1G_ARTSA (SEQ ID NO 34) KSIPYFWEKFDKENYSIWYSEYKYQDELAKVYMSCNLITGMFQRIEKMRKQAFASVCVFG


GTT1_MOUSE (SEQ ID NO 35) ------------------------------------------------------------
GTT1_RAT   (SEQ ID NO 36) ------------------------------------------------------------
EF1G_ARTSA (SEQ ID NO 37) EDNDSSISGIWVWRGQDLAFKLSPDWQIDYESYDWKKLDPDAQETKDLVTQYFTWTGTDK


GTT1_MOUSE (SEQ ID NO 38) ------------
GTT1_RAT   (SEQ ID NO 39) ------------
EF1G_ARTSA (SEQ ID NO 40) QGRKFNQGKIFK
```

Please amend Table 3 as it appears on page 29 as follows:

Table 3

```
EF1G_CAEEL (100-243) (SEQ ID NO 41)----NFD---KKTVEQYK--NELNGQLQVLDRVLVKKTYLVGERLSLADVSVALDLLPAF
SYEP_HUMAN (1-180)   (SEQ ID NO 42)MEHTEIDHWLEFSATKLSSCDSFTSTINELNHCLSLRTYLVGNSLSLADLCVWATLKGNA
                                    ::*   : :. : .  :.:.. :: *:: *  :*****: *****:.*   *

EF1G_CAEEL (100-243) (SEQ ID NO 43)QYVLDANARKSIVNVTRWFRTVVNQPAVKEV--LGEVSLASS-VA-QFNQ--AKFTELS-
SYEP_HUMAN (1-180)   (SEQ ID NO 44)AWQEQLKQKKAPVHVKRWFGFLEAQQAFQSVGTKWDVSTTKARVAPEKKQDVGKFVELPG
                                  :  : : :*: *:*.***  :  * *.:.*    :** :.: ** : :*  .**.**.

EF1G_CAEEL (100-243) (SEQ ID NO 45)---AKVAKSAPKAEKPKKEAKPAAAA--AQP-----E-------DD-EPKEEKS-KDP--
SYEP_HUMAN (1-180)   (SEQ ID NO 46)AEMGKVTVRFPPEASGYLHIGHAKAALLNQHYQVNFKGKLIMRFDDTNPEKEKEDFEKVI
                                   .**:   *   .   .   * **   *      :       ** :*::**.  :
```

Please amend Table 4 as it appears on page 30 as follows:

Table 4

```
EF1G_RABIT (SEQ ID NO 47) MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA
GTH4_MAIZE (SEQ ID NO 48) -ATPAVKVYGWAISPFVSRALLALEEAGVDYELVPMSRQDGD-HRRPEHLARNPFGKVPV
                           *: :: .*     .* :      . * *    .*   : *: :* **.* : * ****.

EF1G_RABIT (SEQ ID NO 49) FEGDDGFCVFESNAIAYYVS----NEELRGSTPEAAAQVVQWVSFADSDIVPPAST----
GTH4_MAIZE (SEQ ID NO 50) LE-DGDLTLFESRAIARHVLRKHKPELLGGGRLEQTAMVDVWLEVEAHQLSPPAIAIVVE
                          :* *..: :***.*** :*      * * *.  * :* *  *:..   :: *** :

EF1G_RABIT (SEQ ID NO 51) WVFPTLGIMHHNKQATENAKEEVKRILGLLDAHLKTRTFLVGERVTLADITVVCTLLWLY
GTH4_MAIZE (SEQ ID NO 52) CVFAPFLGRERNQAVVDENVEKLKKVLEVYEARLATCTYLAGDFLSLADLSPF-TIMHCL
                           **..:   .:*: ..::  *::*::* : :*:* * *:*.*: ::***:: . *::

EF1G_RABIT (SEQ ID NO 53) KQVLEPSFRQAFPNTNRWFLTCINQPQFRAVLGEVKLCEKMAQFDAKKFAESQPKKDTPR
GTH4_MAIZE (SEQ ID NO 54) MATEYAALVHALPHVSAWWQGLAARP---AAN-------KVAQF--MPVGAGAPKEQE--
                            .  .:: :*:*:.. *:     :*   *.        *:***    .. . **::
```

Please amend the paragraph as it appears on page 31, lines 10-24 as follows:

The collection of 804 GPCR sequences and fragments contained several classes (e.g. rhodopsin-like, secretin-like, pheromone, etc.) of proteins. In turn, each of these classes comprised several representatives. Instead of selecting representatives from each of the identified classes, the order of the sequences in this set of 804 members were randomized. Note that the contents of the [sequence] sequences themselves remained unchanged, only their order of appearance was modified. For example, the 613-th sequence was now listed 4-th, the 11-th sequence now appeared in the 45-th position, and so on. Subsequently, a training set T was formed by collecting the sequences and fragments listed in the

first 80 positions, arguably a very small set if one considers the diversity of the GPCR family. Essentially, slightly less than 1/10-th of the available dataset were randomly [sud-selected]sub-selected for the purposes of building the composite descriptor. Table 5 below contains a listing of the labels of the 80 sequences in this training set. Table 5 shows the Swiss-Prot labels of the 80 sequences in the training set for the G protein-coupled receptor experiment. The labels are listed in the order they were selected and they correspond to both sequences and sequence fragments.

Please amend the paragraph as it appears on page 43, lines 10-14 as follows:

The three composite descriptors were used to search the collection of 19,099 ORFs that were reported for the C. elegans genome, by the Washington University in St. Louis, School of Medicine, Genome Sequence Center, [(see: http://genome.wustl.edu/gsc/C_elegans)] as of June 13, 1999. In all three cases, the corresponding values of $Thres_{rand}$ that were established by searching RAND-Swiss-Prot were used.

Please amend the paragraph as it appears on page 48, lines 4-20 as follows:

The fragments were:

```
>Y94H6A_142.g fragment (SEQ ID NO 55)
IFDNTNDLVASLLGISSITVYRKRKRIGEE
>C16C2.1 fragment (SEQ ID NO 56)
YLSGSTRAKLAESLGLSDNQVKVWFQNRRT
>F18C5.2 fragment (SEQ ID NO 57)
ISRSTAKEVATARGISEGTVYSYLAMAVEK
>Y39F10A.a fragment (SEQ ID NO 58)
LSAYTISDLAKHFNVSKIEILKIDIEGAEL
>Y48C3A.s fragment (SEQ ID NO 59)
NEVLNLNEVAKELNISKRRVYDVINVLEGL
```

and their respective top-scoring sequences from the training set of 70 helix-turn helix segments, blast scores, P and N values are:

| # | C. elegans ORF | Top Scoring | Scor | P | N |
|---|---|---|---|---|---|
| 1 | Y94H6A_142.g | RPSF_BACSU | 50 | 2.80E-06 | 1 |
| 2 | C16C2.1 | TER3_ECOLI | 45 | 1.30E-05 | 1 |
| 3 | F18C5.2 | VBP_BPMU | 47 | 9.30E-06 | 1 |
| 4 | Y39F10A.a | TNP0_ECOLI | 39 | 1.10E-04 | 1 |
| 5 | Y48C3A.s | TNP1_ECOLI | 49 | 6.40E-06 | 1 |

IN THE CLAIMS:

Please amend the claims as follows:

4. (Amended) The method of claim 1, wherein the step of discovering is performed without using any knowledge about properties or features of sequences in the set of unaligned sequences.

5. (Amended) The method of claim 1, further comprising the steps of:

if the candidate sequence comprises the predetermined number of patterns, adding the candidate sequence to the set of sequences to create a new set of sequences; and

performing the step of discovering on the new set of sequences.

6. (Amended) The method of claim 1, wherein each sequence comprises a series of symbols and wherein each pattern comprises a plurality of positions, some of the plurality of positions each [comprising]comprise at least one expected symbol and other of the plurality of positions [comprising]comprise ["don't care"] positions which may be occupied by any sequence character.